Méthodes d'alignement des propositions : un défi aux traductions croisées

Yayoi Nakamura-Delloye

TALN 2007, Toulouse, 12–15 juin 2007

Méthodes d'alignement des propositions : un défi aux traductions croisées

Yayoi Nakamura-Delloye
Université Paris 7 – LATTICE
1 rue Maurice Arnoux 92120 Montrouge
http://www.lattice.cnrs.fr
yayoi@free.fr

Résumé Le présent article décrit deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Les deux méthodes sont caractérisées par leur capacité d'alignement des traductions croisées, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases. Contrairement aux résultats obtenus avec l'approche spectrale qui nous paraissent non satisfaisants, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique supporte bien les traductions croisées.

Abstract The present paper describes two methods for clauses alignment. The first one uses a graph matching approach, while the second one relies on agglomerative hirerarchical clustering (AHC). Both methods are characterized by the fact they can align cross translations, which was impossible for previous classic sentence alignment methods. Though the results given by the spectral method are unsatisfactory, the method based on AHC is very promising. It handles correctly cross translations.

Mots-clefs : alignement des corpus parallèles, appariement de graphes, classification ascendante hiérarchique, proposition syntaxique, mémoire de traduction, linguistique contrastive.

Keywords: parallel corpora alignment, graph matching, agglomerative hierarchical clustering, syntactic clause, translation memory, contrastive linguistics.

1 Introduction

L'alignement des propositions consiste en la mise en correspondance des propositions syntaxiques avec leurs traductions dans des textes parallèles. Les corpus parallèles alignés au niveau des propositions pourraient être des ressources profitables dans beaucoup de domaines tels que la traduction automatique ou la traduction assistée par ordinateur (TAO), ainsi que pour les recherches en linguistique contrastive. En dépit de cet intérêt notable, peu de travaux sur l'alignement des propositions ont été réalisés et les quelques méthodes proposées sont semblables à celles pour l'appariement des phrases. Or l'alignement des propositions entre deux langues

relativement différentes sur tous les plans, telles que la paire français-japonais, n'est pas réalisable par la simple application d'une méthode d'alignement des phrases. Nous avons donc essayé de réaliser un système supportant ces différences structurales des langues traitées. Nous décrivons, dans le présent article, deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Nous allons d'abord présenter un bref état de l'art afin de mettre en relief les problèmes. Puis, l'exposé se poursuivra par la description des deux méthodes, pour se terminer par l'analyse des résultats obtenus et une discussion sur les pistes d'amélioration.

2 Problèmes et solution adoptée

Il existe très peu de travaux sur l'alignement des propositions. Nous pouvons tout de même citer ceux de Piperidis, Papageorgiou et Boutsis (BOUTSIS & PIPERIDIS, 1998) (PIPERIDIS *et al.*, 2000) sur les textes parallèles anglais-grec et ceux de Wang et Ren (WANG & REN, 2005) sur le japonais-chinois. Dans la méthode de Piperidis, l'algorithme d'alignement est semblable à celui proposé par Brown, Lai et Mercer (BROWN *et al.*, 1991) pour l'alignement des phrases, à l'exception du fait qu'il utilise des informations lexicales contrairement à la méthode d'alignement des phrases n'exploitant, elle, que la longueur des textes. Wang et Ren améliorent également l'appariement basé sur les longueurs des textes par l'introduction d'un calcul de similarité basé sur l'information portée par les idéogrammes Han.

Il n'existe à ce jour a priori aucune étude sur l'alignement des propositions traitant le japonais, avec le français, ou même avec l'anglais. Il existe cependant un article portant sur l'alignement manuel des propositions anglais-japonais qui présente une méthode d'alignement manuel et les difficultés rencontrées (KASHIOKA *et al.*, 2003).

2.1 Difficultés d'appariement des propositions dues aux différences entre les langues

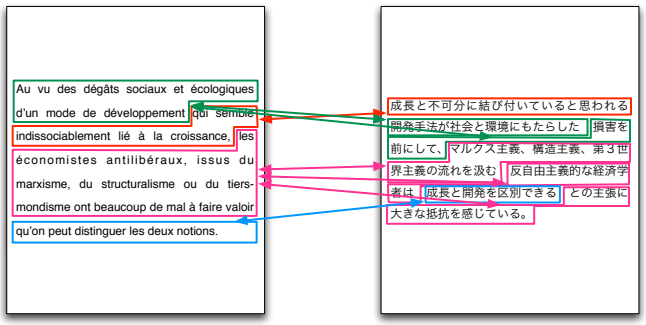

FIG. 1 – Exemple de non-parallélisme de l'alignement des propositions Français-Japonais

Dans leur article, Kashioka *et al.* présentent la constitution d'un corpus parallèle avec alignement au niveau des propositions, réalisée dans un but d'utilisation pour la traduction automatique des monologues (e.g. nouvelles télévisées, conférences, présentations techniques). Une re-

marque intéressante faite par les auteurs suite à cette expérience, porte sur la différence d'ordre des propositions japonaises et des segments anglais correspondants : on constate beaucoup de croisement des alignements. Ce non-parallélisme des propositions (cf. figure 1) implique l'impossibilité d'appliquer les méthodes d'alignement des phrases classiques basées sur l'hypothèse de parallélisme. Nous avons donc besoin, pour automatiser la tâche d'alignement des propositions, de concevoir un autre algorithme qui ne présuppose pas le parallélisme.

2.2 Éléments de solution

Pour supporter les croisements des traductions, l'automatisation de l'opération d'alignement nécessite un algorithme utilisant une structure non linéaire mais à deux dimensions telle que les graphes. Notre idée est comme suit : à l'aide des informations sur les relations entre les propositions, nous construisons l'arbre dépendanciel en propositions (arbre des propositions, ci-après) pour réaliser l'alignement par une méthode d'appariement des graphes (cf. figure 2).

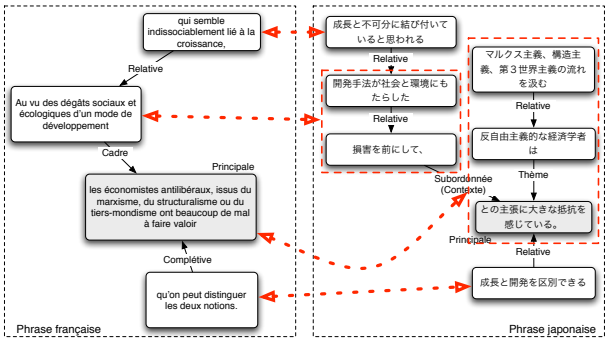

FIG. 2 – Alignement des propositions à l'aide des arbres des propositions

L'alignement à l'aide de graphes n'est pas un concept nouveau. Plusieurs études sur l'alignement anglais-japonais de structures inférieures à la proposition utilisant les arbres syntaxiques ont été réalisées. Les travaux de Matsumoto et *al.* (MATSUMOTO *et al.*, 1993) proposent une méthode permettant de trouver des correspondances structurelles entre deux arbres de dépendance. Dans les méthodes (KAJI *et al.*, 1993) (IMAMURA, 2000) (WATANABE *et al.*, 2000), l'alignement des syntagmes est réalisé sur la base des mots mis en correspondance à l'aide d'un dictionnaire bilingue. Les mots alignés servent à ancrer les textes pour repérer les segments à extraire et la représentation arborescente permet de déterminer correctement les structures formées par ces mots ancres.

Notre approche est semblable à celle de Matsumoto. La difficulté est que la recherche de la meilleure décomposition des arbres pour obtenir les structures isomorphes permettant l'appariement maximal revient à un appariement *many-to-many* de graphes, qui est un problème de grande complexité algorithmique. Dans les travaux de Matsumoto, est retenue une stratégie d'amélioration par l'utilisation de la méthode du *branch-and-bound*. Dans le cadre des présents travaux, nous avons choisi une solution basée sur une technique d'appariement des graphes. En effet, dans la théorie des graphes, il existe un ensemble de méthodes beaucoup plus économiques que les procédures de recherche combinatoire, généralement connues sous le nom de méthodes spectrales.

Néanmoins, cette méthode s'appuie essentiellement sur la topologie des graphes à apparier et

n'est pas destinée à exploiter différentes informations disponibles, notamment les informations lexicales dans le cas de nos travaux. La dernière étape de la méthode spectrale consistant en un regroupement des points projetés, nous a inspiré l'approche pour l'alignement par la classification ascendante hiérarchique (CAH). Celle-ci devant permettre de mieux profiter des informations lexicales tout en supportant les croisements des traductions.

Après examen de l'existant, nous avons réalisé deux méthodes d'alignement des propositions. L'une est basée sur les méthodes d'appariement des graphes – profitant pleinement des structures des arbres des propositions –, l'autre exploitant les informations lexicales et les longueurs tout en supportant les croisements de correspondance avec une méthode inspirée de la classification ascendante hiérarchique.

3 Méthodes basées sur l'approche spectrale

Dans la théorie des graphes, l'appariement par une approche spectrale consiste à représenter et distinguer les propriétés structurales des graphes à l'aide des valeurs propres et des vecteurs propres de leurs matrices d'adjacence et se base généralement sur une technique de décomposition spectrale. L'algorithme sur lequel nous nous sommes plus particulièrement appuyés, celui proposé par Kosinov et Caelli (KOSINOV & CAELLI, 2002) (KOSINOV & CAELLI, 2004), est une amélioration visant en particulier la réalisation des appariements de graphes inexacts. Lerallut (LERALLUT, 2006) a ensuite amélioré cette méthode pour prendre en compte des informations supplémentaires en cas d'appariement de graphes valués.

Dans le cadre de notre alignement des propositions, la méthode de Kosinov est directement utilisée pour apparier les arbres des propositions. Afin d'exploiter au mieux les informations disponibles pour réaliser un meilleur appariement, nous avons également réalisé une adaptation de la méthode de Lerallut à notre opération d'alignement des propositions.

3.1 Méthodes spectrales pour l'appariement des graphes

La méthode d'appariement de graphes inexact proposée par Kosinov et Caelli combine les avantages des techniques de décomposition spectrale, de projection et de classification (*clustering*). Elle consiste, étant donné les matrices d'adjacence A_1 et A_2 créées à partir des graphes G_1 et G_2 respectivement, (i) à calculer les valeurs propres et les vecteurs propres, (ii) à tronquer les matrices selon le nombre de dimensions choisi pour la projection, (iii) à normaliser les valeurs propres et les vecteurs propres pour projeter ensuite chaque graphe, enfin (iv) à réaliser l'appariement par regroupement des nœuds projetés à l'aide d'un algorithme de classification.

Après la décomposition, les données relatives aux nœuds obtenues avec la matrice d'adjacence sont projetées sur les k vecteurs propres les plus importants, formant un sous-espace propre de dimension réduite du graphe. Dans ce sous-espace propre, des nœuds ou des ensembles de nœuds ayant des propriétés structurales semblables sont proches les uns des autres, permettant ainsi une comparaison et un appariement des graphes. Néanmoins, étant donné que les graphes à aligner peuvent posséder un nombre différent de nœuds, une opération de normalisation est également nécessaire pour assurer de bonnes conditions de comparaison. Par examen du positionnement de ces projections de nœuds, la mise en correspondance est alors possible. Le regroupement des points projetés par une méthode de classification ascendante hiérarchique

permet de réaliser l'appariement des ensembles de nœuds entre les graphes.

Lerallut, cherchant à appliquer cette méthode à un traitement des images, propose une amélioration permettant de prendre en compte des informations supplémentaires en cas d'appariement de graphes valués. Sa méthode part du résultat obtenu avec la méthode de Kosinov. Cette dernière permet d'abord d'obtenir la matrice topologique contenant des distances euclidiennes entre toutes les projections dans le sous-espace propre. Les graphes sont ensuite valués par l'affectation de couleurs à chaque nœud, et une matrice des distances de couleurs est calculée entre tous les nœuds des deux graphes. Après avoir normalisé ces deux matrices en les divisant par leur valeur maximum, on calcule une somme pondérée. Enfin, après une modification pour écarter les valeurs très distantes, un sous-espace propre de cette matrice est calculé afin d'y projeter tous les nœuds.

3.2 Application de la méthode spectrale à l'alignement des propositions

L'alignement des propositions réalisé par la méthode de Kosinov s'appuie uniquement sur la topologie des graphes. Toutefois, les arbres des propositions dont nous disposons comme entrée du système contiennent beaucoup plus d'informations qui pourraient être utilisées au profit d'un bon appariement. Afin d'exploiter au mieux ces informations disponibles, nous avons tout d'abord tenté d'adapter la méthode de Lerallut de sorte que les graphes à apparier soient valués, non par l'affectation de couleurs, mais selon les types de propositions. Mais, afin de calculer la distance entre deux nœuds sur la base de leur type de proposition, il nous a d'abord fallu définir une distance entre chaque type de proposition.

Faute de corpus *ad. hoc* en quantité suffisante pour le calcul des probabilités des correspondances entre les types, nous avons choisi une méthode plus empirique, qui permettra également de constituer un premier corpus pour des travaux futurs. Nous avons d'abord mis en correspondance les types de propositions du français et du japonais, qui semblaient les plus proches sur le plan syntaxique. Étant donné l'existence d'un lien non négligeable entre les fonctions syntaxiques et la place dans la phrase, nous avons défini une distance entre chaque type de proposition sur la base de la topologie de la phrase. À cette fin, nous nous sommes principalement appuyés sur la structure canonique de la phrase française. La structure canonique est définie comme : éléments extra-prédicatifs – thème – racine (principale) – subordonnées post-nominales – subordonnées post-verbales – subordonnées périphériques[1]. Le principe de base est que la distance d'un type donné de proposition par rapport à la racine, point central, est définie par le nombre de propositions susceptibles d'apparaître entre elles.

Nous avons utilisé les distances ainsi définies pour calculer la matrice des distances de couleurs et réalisé l'appariement des graphes avec la méthode de Lerallut. Mais, nous n'avons pas obtenu le résultat souhaité : l'appariement ne reflétait pas bien les distances des types de propositions. Afin de mieux refléter les informations sur les types de nœuds tout en conservant la structure des arbres d'entrée, nous avons introduit une autre formule. Le principe du nouveau calcul consiste à prendre compte des informations topologiques pour les relations entre les nœuds du même arbre et des informations sur les types pour les distances entre les nœuds d'arbres différents.

Étant donné les deux graphes X et Y, la matrice finale de la méthode de Lerallut est une matrice M_{final} de $|X|+|Y| \times |X|+|Y|$, $M_{\mathit{final}}(i,j)$ correspondant à la somme des distances topologique et de couleur normalisées entre les nœuds i et j.

[1] Pour une description détaillée des types de propositions, voir (NAKAMURA-DELLOYE, 2007).

Nous décomposons cette matrice finale comme :

$$M_{final} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

de manière à obtenir les sous-matrices M_{11} comme une matrice $|X| \times |X|$, M_{12} comme $|X| \times |Y|$, M_{21} comme $|Y| \times |X|$ et M_{22} comme $|Y| \times |Y|$, où :

$$\begin{aligned} M_{11}(i,j) &= \mathrm{dist}_{\mathrm{topo}}(X_i, X_j) \times (1-\alpha) \\ M_{12}(i,j) &= \mathrm{dist}_{\mathrm{type}}(X_i, Y_j) \times \alpha \\ M_{21}(i,j) &= \mathrm{dist}_{\mathrm{topo}}(X_j, Y_i) \times (1-\alpha) \\ M_{22}(i,j) &= \mathrm{dist}_{\mathrm{type}}(Y_i, Y_j) \times \alpha \end{aligned}$$

4 Méthode inspirée de la classification ascendante hiérarchique

La deuxième méthode que nous avons décidé d'étudier est basée sur la classification ascendante hiérarchique (CAH), celle-ci devant permettre de mieux profiter des informations lexicales tout en supportant les croisements des traductions. En effet, nous considérons maintenant l'alignement, comme nous l'avons fait dans la méthode spectrale, comme le regroupement des points semblables appartenant à deux classes différentes.

4.1 Deux matrices de base

Nous créons tout d'abord deux matrices : une contenant les similarités de chaque paire de propositions (M_{sim}), et une autre pour stocker les valeurs indiquant l'évolution du rapport des longueurs entre les propositions de langues différentes ($M_{raplong}$).

Étant donné les deux (ensembles de) phrases X et Y, la matrice de similarité M_{sim} de $(|X| + |Y|) \times (|X| + |Y|)$ est définie comme :

$$M_{sim} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

avec les sous-matrices M_{11} de $|X| \times |X|$, M_{12} de $|X| \times |Y|$, M_{21} de $|Y| \times |X|$ et M_{22} de $|Y| \times |Y|$, où :

$$\begin{aligned} M_{11}(i,j) &= synt(X_i, X_j) \\ M_{12}(i,j) &= simlex(X_i, Y_j) \\ M_{21}(i,j) &= simlex(X_j, Y_i) \\ M_{22}(i,j) &= synt(Y_i, Y_j) \end{aligned}$$

$simlex(X_i, Y_j)$ est la similarité lexicale obtenue de manière classique telle qu'avec le coefficient de Dice. Dans notre réalisation, elle est calculée à l'aide d'un dictionnaire bilingue et d'une liste de mots alignés au moment de l'alignement des phrases du même corpus. Lorsque la similarité lexicale est à 0, on lui donne la valeur minimum α pour favoriser la fusion des éléments (propositions) appartenant aux classes différentes. $synt(X_i, X_j)$ est obtenue de la même manière qu'une matrice d'adjacence, c'est-à-dire 0 s'il n'existe aucun arc entre les nœuds i et j dans l'arbre d'entrée, et β s'il en existe un. Ce mécanisme permet, dans le cas du regroupement d'éléments appartenant à la même classe, de réaliser l'agrégation entre deux éléments en relation de dépendance, plutôt qu'entre deux éléments qui n'en ont aucune.

La matrice d'évolution du rapport des longueurs $M_{raplong}$ est définie telle qu'à chacun de ses éléments (i, j) corresponde l'évolution du rapport des longueurs entre les propositions de langues différentes, qui se produira si le regroupement des deux éléments considérés, i et j, a lieu. La valeur indiquant cette évolution est pondérée par a afin de pénaliser les fusions importantes d'éléments.

$$M_{raplong}(i,j) = \left(rap\left(F(i,j), J(i,j)\right) - \min\left(rap\left(F(i), J(i)\right), rap\left(F(j), J(j)\right)\right)\right) \cdot a$$

où
- $rap(x, y)$ est le rapport des longueurs des éléments (ou des classes) x et y ;
- $F(x)$ (resp. $J(x)$) est la longueur normalisée de (l'ensemble des) proposition(s) française(s) (resp. japonaise(s)), constituant l'élément (ou la classe) x ;
- $F(x, y)$ (resp. $J(x, y)$) est la longueur normalisée de l'ensemble des propositions françaises (resp. japonaises) constituant la classe regroupant les éléments (ou les classes) x et y ;
- a est le poids défini comme le logarithme de la moyenne des deux longueurs normalisées $a = \log\left(\frac{F(i,j)+J(i,j)}{2}\right)$.

4.2 Agrégation et recalcul des matrices

En combinant ces deux matrices, de similarité et d'évolution du rapport des longueurs, une troisième matrice, matrice courante, est calculée et recalculée après chaque agrégation de deux éléments. La matrice courante est définie comme :

$$M_{courante}(i,j) = \frac{M_{raplong}(i,j)}{M_{sim}(i,j)}$$

Dans cette matrice courante, nous cherchons la valeur minimum pour réaliser l'agrégation de deux éléments (ou classes). Après l'agrégation des deux éléments, nous recalculons la matrice d'évolution du rapport des longueurs, en tenant compte du changement de longueurs des éléments regroupés. La matrice de similarité est également recalculée selon le critère d'agrégation adopté. Dans notre réalisation, les similarités des classes nouvellement créées suite à l'agrégation sont obtenues en divisant la somme des similarités des éléments (ou classes) regroupés, par la valeur v calculée sur la base du nombre de propositions faisant partie de cette nouvelle classe. À partir de ces deux matrices nouvellement calculées, on calcule à nouveau la matrice courante et recommençons les opérations d'agrégation tout comme la CAH. À la différence de l'algorithme classique de CAH, l'itération s'arrête dans notre opération dès que toutes les propositions sont regroupées avec au moins une proposition de l'autre langue.

5 Évaluation des méthodes

Nous avons réalisé une évaluation des méthodes proposées avec quatre corpus parallèles de natures diverses et de langues originaires différentes (1, 2 en français et 3, 4 en japonais) : (1) corpus LMD constitué d'articles du Monde Diplomatique, (2) corpus BRVF et (3) BRVJ composés respectivement de deux et d'un brevets techniques, (4) corpus FdT, un extrait du roman « La fin des temps » de Haruki MURAKAMI. Le corpus est d'abord aligné au niveau des phrases par notre système d'alignement des phrases (NAKAMURA-DELLOYE, 2005) et le résultat est vérifié manuellement. Puis, pour chaque phrase, la détection des propositions

est réalisée à l'aide de nos détecteurs de propositions du français (NAKAMURA-DELLOYE, 2006) et du japonais, et le résultat d'analyse est également corrigé manuellement.

	Caractéristiques				Résultat								
	(A/B)	(C)	(D)	(E)	Partiel (F)			Exact (G)			Paires créées (H)		
	Phr.	Fr	Jp	Prop.	M1	M2	M3	M1	M2	M3	M1	M2	M3
LMD	222/500	644	1026	583	0,643	0,784	0,951	0,127	0,200	0,591	0,813	0,746	0,918
BRVF	161/339	447	854	444	0,619	0,705	0,977	0,081	0,158	0,706	0,750	0,757	0,867
BRVJ	44/66	146	280	141	0,663	0,689	0,990	0,048	0,078	0,537	0,738	0,638	0,674
FdT	99/200	286	428	251	0,670	0,659	0,932	0,138	0,151	0,464	0,892	0,817	0,936

TAB. 1 – Résultats de l'alignement par les trois méthodes

Nous avons utilisé au total 1105 paires de phrases alignées (détails pour chaque corpus indiqués en (B), tableau 1). Parmi celles-ci, nous n'avons pris en compte dans nos résultats d'évaluation que les paires comportant plus d'une proposition dans chaque langue, soit 526 paires de phrases (A), qui représentent 1523 propositions françaises (C) et 2588 propositions japonaises (D), composant 1419 paires de propositions en relation de traduction (E). Nous pouvons constater que le nombre de propositions japonaises est au moins 50% plus élevé que celui des propositions françaises. Cela implique que le modèle de traduction 1-1 (modèle pour la paire en relation de traduction constituée d'une unité dans une langue avec une unité de l'autre langue) est beaucoup moins courant que dans le cas de l'alignement des phrases. En effet, les paires 1-1 représentent moins de 50% et le nombre d'alignements d'une proposition française avec de 2 à plus de 4 propositions japonaises s'élève à environ 40%. Ce type de paire complexe est une source de perturbation pour les méthodes d'alignement des phrases classiques.

Dans le tableau 1, est présenté le résultat de notre évaluation des trois méthodes : méthode spectrale uniquement topologique (M1), méthode spectrale avec types de propositions (M2) et méthodes avec classification ascendante hiérarchique (M3). La zone marquée « Partiel » (F) indique la proportion de paires partiellement correctes parmi l'ensemble des paires effectivement alignées, et la zone marquée « Exact » (G), celle des paires exactement alignées correctement. Enfin, la zone marquée « Paires créées » (H), correspond à la proportion du nombre de paires créées par rapport au nombre correct de paires.

Les résultats montrent que nous avons réussi à améliorer la méthode de Kosinov (M1) avec l'introduction des types de proposition (M2), encore que les chiffres obtenus ne soit pas encore satisfaisants. En effet, il existe beaucoup de cas où la topologie des graphes n'est pas suffisante pour l'appariement des arbres syntaxiques : il arrive, notamment, qu'un arbre soit interprété comme symétrique, alors qu'il ne l'est pas. C'est par exemple le cas des arbres des propositions (présentés figure 3) des phrases parallèles suivantes :

> **Phrase française :** F1 : RACINE Paris avait estimé, à l'époque || F2 : SUBQ , qu'une référence aux valeurs religieuses n'était pas acceptable || F3 : SUBP car elle soulevait des problèmes politiques et constitutionnels en France.
>
> **Phrase japonaise :** J1 : EXTRA *tôji,*(à l'époque) || J2 : THEME *furansu wa,* (La France) || J3 : THEME *shûkyôteki kachi eno genkyû wa* (une référence aux valeurs religieuses) || J4 : SUBAGG *koku-nai de seijijô, kenpôjô no mondai wo hikiokosuga yueni* (car [elle] soulève des problèmes politiques et constitutionnels dans le pays) || J5 : SUBCIT *mitomerarenai tono* (qui dit que ce n'était pas acceptable) || J6 : RACINE *shisei wo totta.*([La France] a pris la position)

Dans les cas comme cet exemple, l'introduction des informations sur le type de proposition a permis d'améliorer le résultat et de fournir un alignement correct.

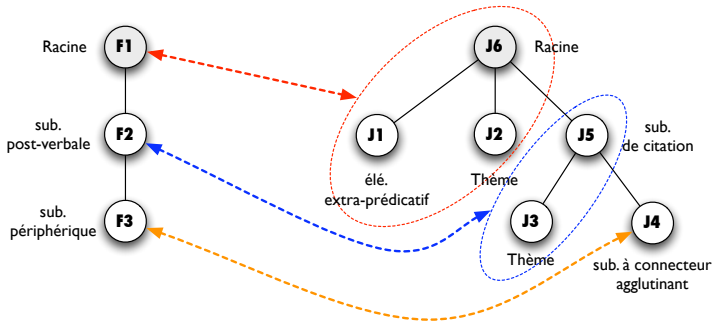

FIG. 3 – Arbres des propositions et appariement correct de leurs nœuds

Toutefois, beaucoup de phrases nécessitaient encore plus d'informations et leur alignement n'a été amélioré qu'avec la méthode à classification ascendante hiérarchique (M3) basée sur la similarité lexicale. Cette méthode possède encore des points potentiels d'amélioration (comme la désambiguïsation lexicale par exemple), mais la capacité d'alignement avec des croisements est un atout crucial. De plus, comme nous le savons bien, la méthode de classification nous permet de définir nous-même la fin du développement des fusions. Par ce mécanisme, nous pourrions obtenir un résultat moins robuste mais plus fiable.

6 Conclusion et perspectives des travaux futurs

Nous avons présenté deux méthodes pour l'alignement des propositions des textes parallèles français-japonais. L'une s'appuie sur une méthode d'appariement des graphes consistant à projeter les nœuds sur un sous-espace propre. L'autre est basée sur une méthode inspirée de la classification ascendante hiérarchique.

Les deux méthodes sont caractérisées par leur capacité d'alignement des traductions croisées dans l'ordre d'apparition, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases. Le résultat obtenu avec la méthode spectrale n'était pas satisfaisant. Il est en effet difficile de trouver une formule permettant de refléter les informations supplémentaires autres que la topologie. Une très récente étude (FRAIKIN *et al.*, 2006) propose une amélioration visant le traitement des graphes orientés. Néanmoins, du fait des différences considérables de structures, l'application de cette méthode à l'alignement des langues très différentes semble difficile. En revanche, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique permet d'exploiter plus efficacement différentes informations tout en supportant les croisement des traductions.

À travers cette expérience, nous avons également rencontré beaucoup de constructions pour lesquelles un appariement même manuel était très difficile. Ces exemples sont, pour nous, non seulement des indicateurs de futurs obstacles à franchir, mais aussi très enrichissants du point de vue de l'étude contrastive sur les structures syntaxiques des phrases française et japonaise. Nos travaux sur l'alignement pourraient non seulement participer au développement du domaine du TAL, mais aussi contribuer aux progrès des études linguistiques contrastives du français-japonais, qui favoriseraient à leur tour l'innovation de nos recherches.

Références

BOUTSIS S. & PIPERIDIS S. (1998). Aligning clauses in parallel texts. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, p. 17 – 26.

BROWN P. F., LAI J. C. & MERCER R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, p. 169 – 176.

FRAIKIN C., NESTEROV Y. & P. V. D. (2006). A gradient-type algorithm optimizing the coupling between matrices and application to graph matching. In *Proceedings of the 13-th ILAS conference in Amsterdam*.

IMAMURA K. (2000). A hierarchical phrase alignment from english and japanese bilingual text. In *Proceedings of CICLing 2001*.

KAJI H., KIDA Y. & MORIMOTO Y. (1993). Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, p. 672–678.

KASHIOKA H., MARUYAMA T. & TANAKA H. (2003). Building a parallel corpus for monologue with clause alignment. In *Proceedings of the 9th Machine Translation Summit*, p. 216 – 223.

KOSINOV S. & CAELLI T. (2002). Inexact multisubgraph matching using graph eigenspace and clustering models. In *Proceedings of SSPR/SPR*, volume 2396, p. 133–142.

KOSINOV S. & CAELLI T. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE transactions on pattern analysis and machine intelligence*, **26**(4), 515–519.

LERALLUT R. (2006). *Modélisation et interprétation d'images à l'aide de graphes*. Thèse de doctorat, École des Mines de Paris.

MATSUMOTO Y., ISHIMOTO H. & UTSURO T. (1993). Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, p. 23– 30.

NAKAMURA-DELLOYE Y. (2005). Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais. In *RECITAL 2005*.

NAKAMURA-DELLOYE Y. (2006). Détection automatique des propositions syntaxiques du français. In *TALN 2006*.

NAKAMURA-DELLOYE Y. (2007). Typologie des subordonnées et des connecteurs en vue de la détection automatique des propositions syntaxiques du français. In *Description linguistique pour le traitement automatique du français*, Cahiers du Cental. Presses universitaires de Louvain. (à paraître).

PIPERIDIS S., PAPAGEORGIOU H. & BOUTSIS S. (2000). From sentences to words and clauses. In J. VÉRONIS, Ed., *Parallel text processing*, p. 117 – 138. Kluwer Academic Publishers.

WANG X. & REN F. (2005). Chinese-japanese clause alignment. *Lecture Notes in Computer Science*, **3406**, 400–412.

WATANABE H., KUROHASHI S. & ARAMAKI E. (2000). Finding structural correspondences from bilingual parsel corpus for corpus-based traslation. In *Proceedings of COLING 2000*, p. 906–912.